Wrocław University of Technology

Podstawowe modele probabilistyczne

Maciej Zięba

maciej.zieba@pwr.edu.pl

Rozpoznawanie Obrazów, Lato 2018/2019

Pojęcie prawdopodobieństwa

- Prawdopodobieństwo reprezentuje **częstość występowania** pewnego **zjawiska**.
- Wyrażone jest pewną **wartością rzeczywistą** z przedziału $[0, 1]$.
- **Suma prawdopodobieństw** wszystkich możliwych **zdarzeń** jest równa 1.
- Możliwe **zdarzenia** reprezentowane są za pomocą **zmiennej losowej** X.

Zmienna losowa

- Zmienna losowa to **funkcja** $X(\omega) : \Omega \to \mathcal{X}$, która danemu stanowi świata (zdarzeniu elementarnemu) ω przyporządkowuje pewną wartość x.
- Możemy poznać / zmierzyć **jedynie wartości** tej funkcji, nigdy argumenty.
- Najwięcej, co możemy odkryć, to **rozkład zmiennej losowej** $p(x)$, czyli jak często pojawia się konkretna wartość zmiennej.

Przykład: *rzut kostką sześcienną*

$$\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$$
$$p(x) = 1/6$$

Zmienne niezależne

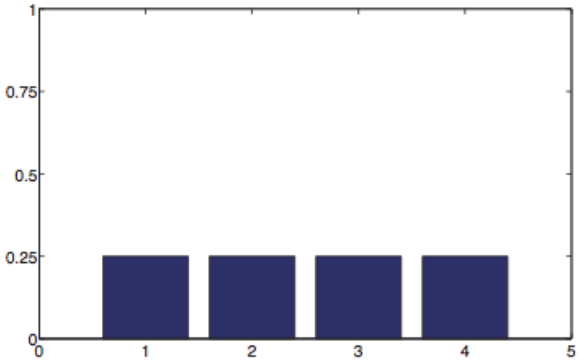
- Mówimy, że dwie zmienne losowe X, Y są **niezależne**, gdy rozkład spełnia warunek

$$p(x, y) = p(x)p(y)$$

- Możemy próbować przewidywać wartość y, obserwując x, jedynie wtedy, gdy **istnieje zależność** pomiędzy nimi.

Dyskretne zmienne losowe

- Zbiór możliwych wartości $\mathcal{X}$ jest **co najwyżej przeliczalny**.



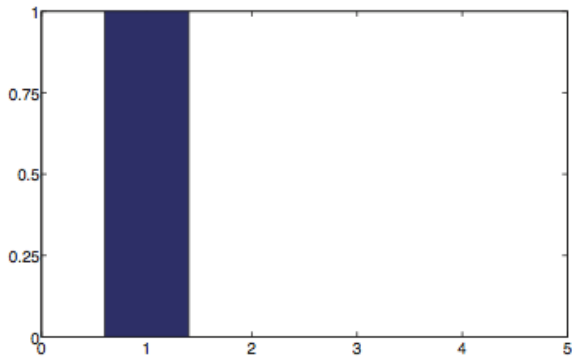
- Suma prawdopodobieństw równa się 1:

$$\sum_{x \in \mathcal{X}} p(X = x) = \sum_{x \in \mathcal{X}} p(x) = 1.$$

- Dla podzbioru $\mathcal{A} \subseteq \mathcal{X}$:

$$\sum_{x \in \mathcal{A}} p(x) = 1 - \sum_{x \in \neg \mathcal{A}} p(x),$$

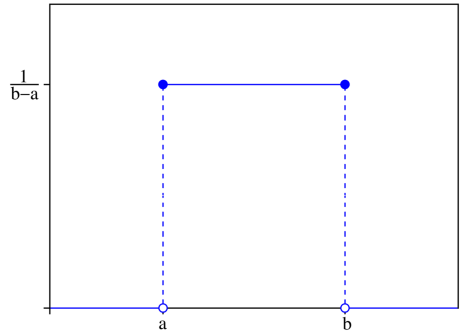
gdzie $\neg \mathcal{A}$ stanowi dopełnienie zbioru $\mathcal{A}$.



Ciągłe zmienne losowe

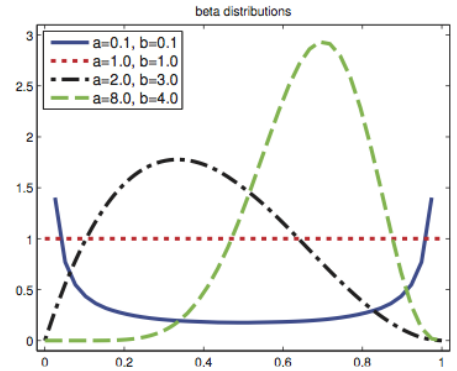
- Rozkład opisuje **funkcją gęstości** $p(x)$.
- Prawdopodobieństwo $X \in [a, b]$:

$$p(a \leqslant X \leqslant b) = \int_a^b p(x)dx$$

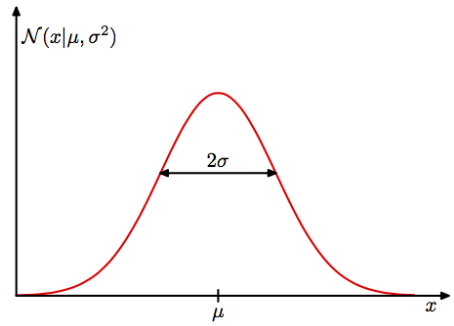


- Dystrybuanta:

$$p(X \leqslant b) = F(b) = \int_{-\infty}^b p(x)dx$$



- Funkcja gęstości **spełnia reguły**:
  brzegową: $p(x) = \int p(x, y)dy$
  łańcuchową: $p(x, y) = p(x|y)p(y)$



Wartość oczekiwana

- **Wartość oczekiwana** jest średnią wartością zmiennej losowej. Dla rozkładu **dyskretnego** definiuje się następująco:

$$E[X] = \sum_{x \in \mathcal{X}} x \cdot p(x),$$

dla rozkładu **ciągłego** definiuje się następująco:

$$E[X] = \int_{x \in \mathcal{X}} x \cdot p(x) dx,$$

- Wartość oczekiwana dla rzutu niesymetryczną monetą:

$$E[X] = 1 \cdot \theta + 0 \cdot (1 - \theta) = \theta.$$

Wariancja i odchylenie standardowe

- **Wariancja** opisuje **rozpiętość** rozkładu (jak bardzo odchylają się wartości x od średniej) i wyraża się następująco:

$$Var[X] = E[(X - E[X])^2] = E[X^2] - (E[X])^2$$

- **Odchylenie standardowe** definiowane jest jako **pierwiastek** z wariancji:

$$std[X] = \sqrt{Var[X]}$$

- Wariancja dla rzutu monetą:

$$Var[X] = \theta \cdot (1 - \theta).$$

Korelacja i kowariancja

- Miarą **liniowej zależności** pomiędzy zmienną losową X i Y jest **kowariancja** zadana wzorem:

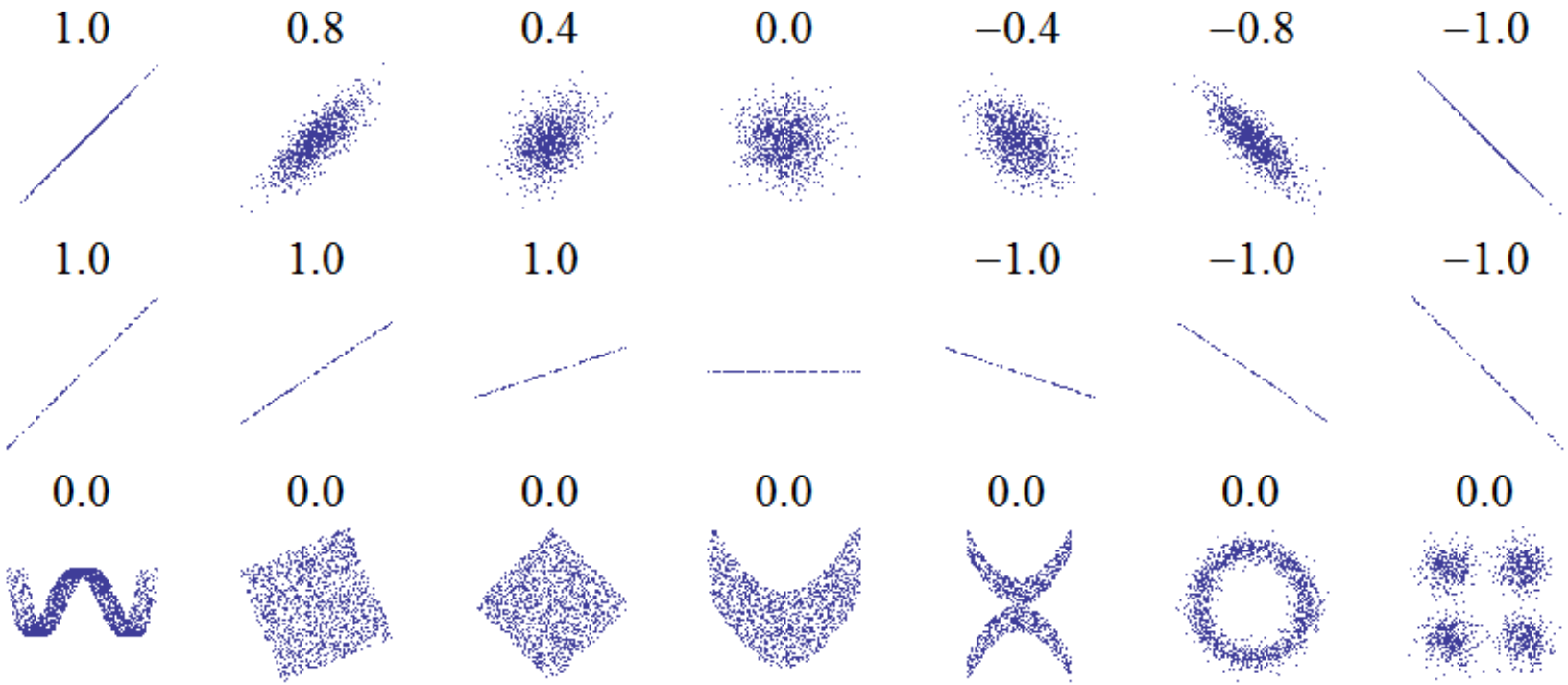
$$\begin{aligned} cov[X,Y] &= E[(X - E[X])(Y - E[Y])] \\ &= E[XY] - E[X]E[Y]. \end{aligned}$$

- **Kowariancja** przyjmuje **wartości z przedziału** $(-\infty, \infty)$, w praktyce wygodniej jest operować na znormalizowanej postaci **kowariancji** nazywanej **korelacją**:

$$corr[X,Y] = \frac{cov[X,Y]}{\sqrt{Var[X]Var[Y]}},$$

która przyjmuje **wartości z przedziału** $[-1, 1]$.

Korelacja i kowariancja



Jeśli dwie zmienne są nieskorelowane, to **nie oznacza**, że są **niezależne**.

Rozkład dwupunktowy

- Zmienna losowa przyjmuje wartości ze zbioru $x \in \{0, 1\}$.
- Funkcja **rozkładu prawdopodobieństwa**:

$$\text{Bern}(x|\theta) = \theta^{\mathbb{I}(x=1)}(1-\theta)^{\mathbb{I}(x=0)} = \theta^x(1-\theta)^{1-x}$$

- Rozkład ma interpretację **pojedynczego rzutu monetą**.
- Parametr θ reprezentuje **prawdopodobieństwo reszki w rzucie monetą**.
- Podstawowe **własności** rozkładu:

$$E[X] = \theta,\ Var[X] = \theta \cdot (1-\theta).$$

Rozkład wielopunktowy

- Wektory binarnych zmiennych losowych o wartościach $x_i \in \{0, 1\}$ spełniających warunek $\sum_{i=1}^{K} x_i = 1$.
- **Rozkład prawdopodobieństwa**:

$$\mathrm{Cat}(\mathbf{x}|\boldsymbol{\theta}) = \prod_{i=1}^{K} \theta_i^{\mathbb{I}(x_i=1)}$$

- Rozkład ma interpretację **rzutu K-ścienną kostką**.
- Parametr θ_i reprezentuje **prawdopodobieństwo wypadnięcia i oczek**.
- Podstawowe **własności** rozkładu:

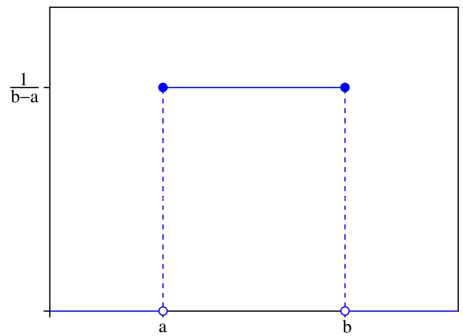
$$E[X_j] = \theta_j,\ Var[X_j] = \theta_j \cdot (1 - \theta_j),\ cov[X_j, X_i] = -\theta_j\theta_i$$

Rozkład jednostajny

- Funkcja **gęstości**, $x \in \mathbb{R}$:

$$\mathcal{U}(x|a,b) = \begin{cases} \frac{1}{b-a}, & x \in [a,b], \\ 0, & \text{w.p.p.} \end{cases}$$

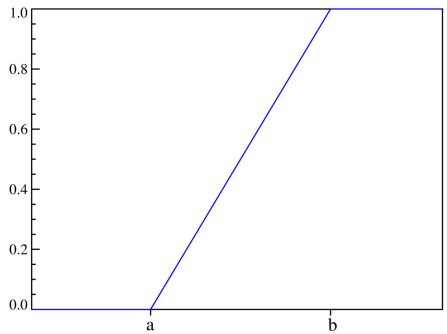
funkcja gęstości



- Podstawowe **własności** rozkładu:

$$E[X] = \frac{a+b}{2},\ Var[X] = \frac{(b-a)^2}{12}.$$

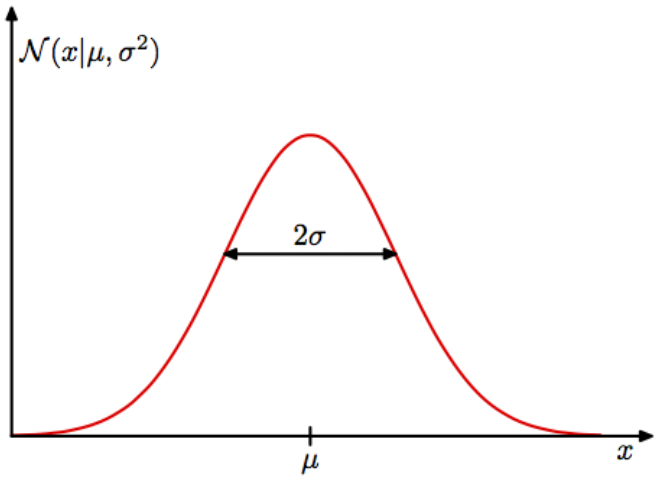
dystrybuanta



Rozkład Gaussa / normalny

- Funkcja **gęstości**, $x \in \mathbb{R}$:

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



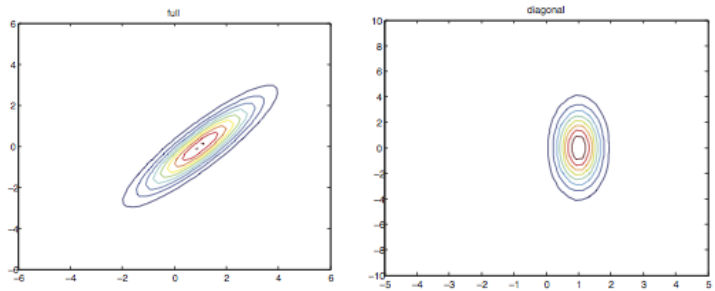
- Podstawowe **własności** rozkładu:

$$E[X] = \mu,\ Var[X] = \sigma^2.$$

Wielowymiarowy rozkład Gaussa / normalny

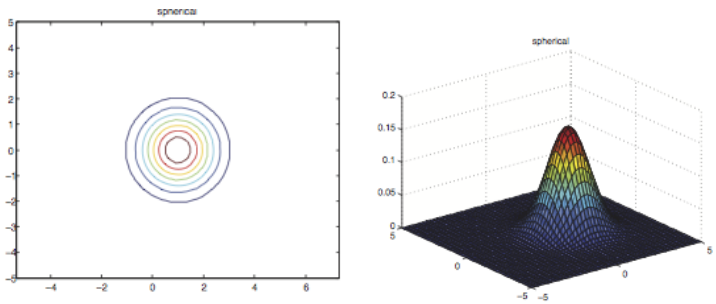
- Funkcja **gęstości**, $\mathbf{x} \in \mathbb{R}^K$:

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{K}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}} \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})}$$



- Podstawowe **własności** rozkładu:

$$E[\mathbf{x}] = \boldsymbol{\mu},\ cov[\mathbf{x}] = \boldsymbol{\Sigma}.$$



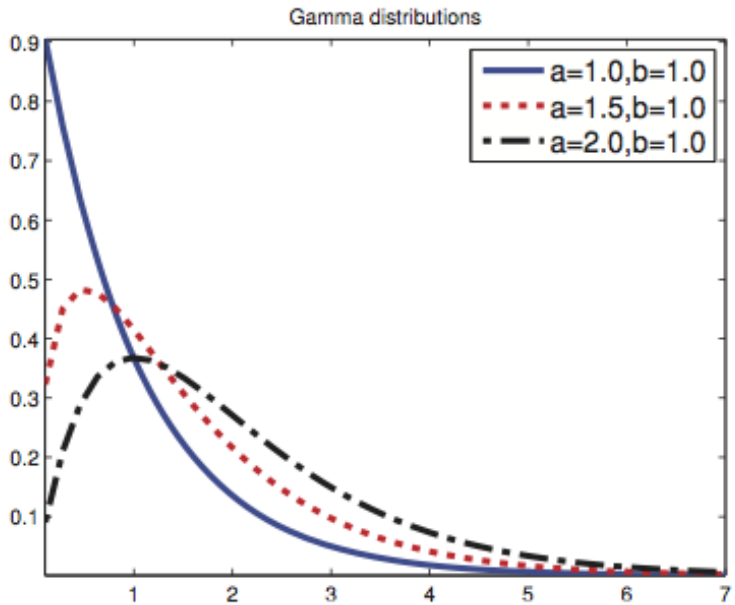
- Inną parametryzacją jest macierz precyzji $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$.

Rozkład Gamma

- Funkcja **gęstości**, $x > 0$:

$$Ga(x|a,b) = \frac{b^a}{\Gamma(b)} x^{a-1} e^{-bx},$$

$$\Gamma(b) = \int_0^\infty u^{b-1} e^{-u} du.$$



- Podstawowe **własności** rozkładu:

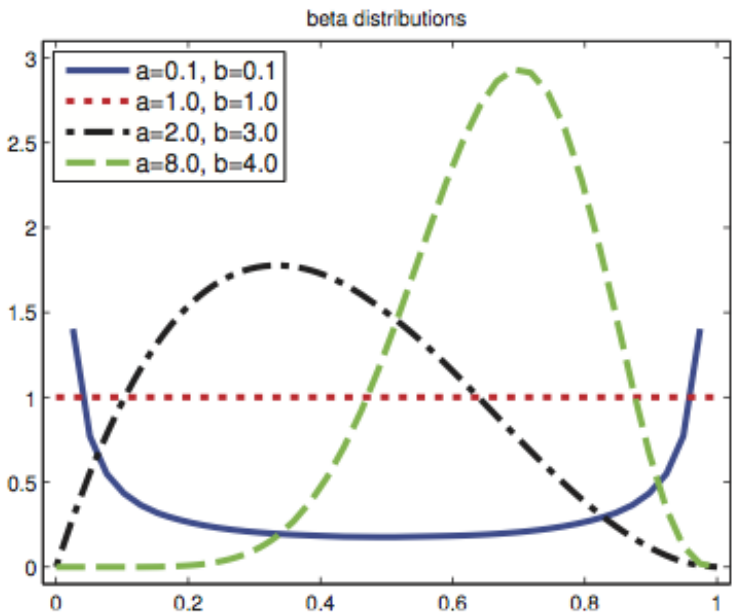
$$E[X] = \frac{a}{b},\ Var[X] = \frac{a}{b^2}.$$

Rozkład Beta

- Funkcja **gęstości**, $x \in [0, 1]$:

$$Beta(x|a, b) = \frac{x^{a-1}(1-x)^{b-1}}{B(a, b)},$$

$$B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$



- Podstawowe **własności** rozkładu:

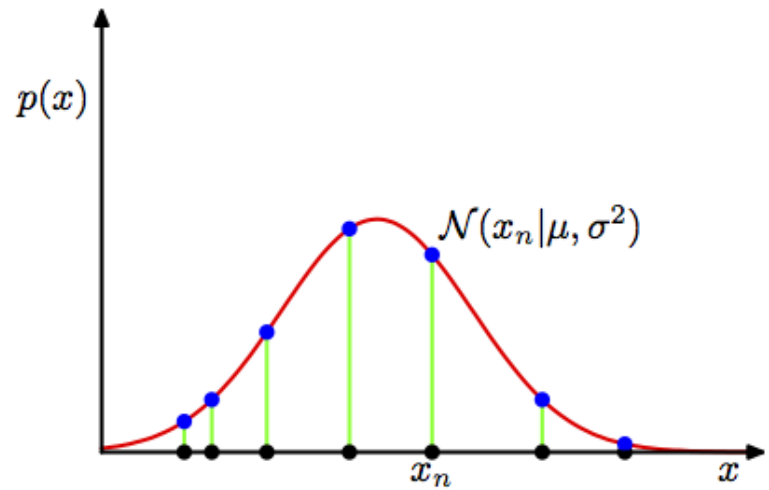
$$E[X] = \frac{a}{a+b}$$

$$Var[X] = \frac{ab}{(a+b)^2(a+b+1)}.$$

Funkcja wiarygodności

- Dysponujemy **niezależnymi obserwacjami** $\mathcal{D} = \{\mathbf{x}_n\}_{n=1}^N$.
- Rozpatrujemy **model** $p(\mathbf{x}|\boldsymbol{\theta})$ dla rozkładu, z którego pochodzą **dane**.
- **Funkcję wiarygodności** (*ang. likelihood function*) definiuje się następująco:

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{n=1}^{N} p(\mathbf{x}_n|\boldsymbol{\theta}).$$

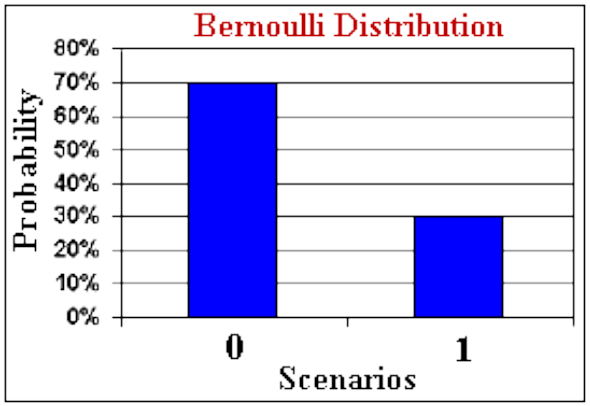


Funkcja wiarygodności - Przykład 1

Dysponujemy ciągiem obserwacji dotyczących rzutu monetą $\mathcal{D} = \{0, 0, 0, 1, 1, 0, 1, 0, 1\}$

Chcemy ocenić, czy bardziej **wiarygodne** jest, że dane zostały **wygenerowane**:

- z rozkładu $\mathrm{Bern}(x|0.5)$,
- czy też z rozkładu $\mathrm{Bern}(x|0.75)$.



Wartości **funkcji wiarygodności** $p(\mathcal{D}|\theta)$ dla rozkładów wynoszą:

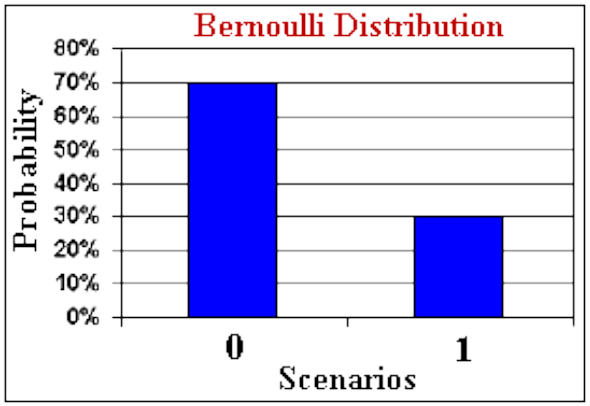
$$p(\mathcal{D}|0.5) = 0.5^9 = 0.001953125,$$
$$p(\mathcal{D}|0.75) = 0.75^4 \cdot 0.25^5 = 0.00030899.$$

Funkcja wiarygodności - Przykład 1

Dysponujemy ciągiem obserwacji dotyczących rzutu monetą $\mathcal{D} = \{0, 0, 0, 1, 1, 0, 1, 0, 1\}$

Chcemy ocenić, czy bardziej **wiarygodne** jest, że dane zostały **wygenerowane**:

- z rozkładu $\mathrm{Bern}(x|0.5)$,
- czy też z rozkładu $\mathrm{Bern}(x|0.75)$.



Wartości **funkcji wiarygodności** $p(\mathcal{D}|\theta)$ dla rozkładów wynoszą:

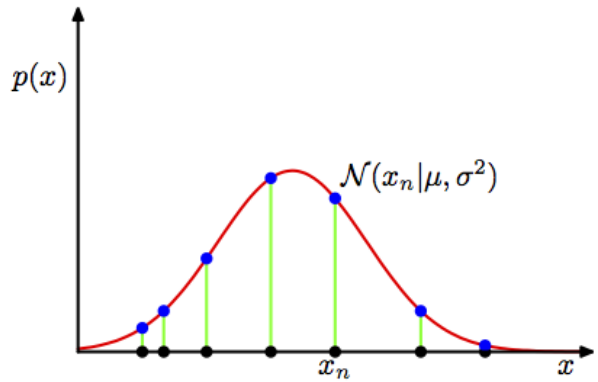
$p(\mathcal{D}|0.5) = 0.5^9 = 0.001953125,$
$p(\mathcal{D}|0.75) = 0.75^4 \cdot 0.25^5 = 0.00030899.$

Funkcja wiarygodności- Przykład 2

Dysponujemy ciągiem obserwacji
$\mathcal{D} = \{0.24, 0.32, 0.21, -0.2, -0.87,$
$0.23, -0.12, 0.01\}$

Chcemy ocenić, czy bardziej **wiarygodne** jest, że dane zostały **wygenerowane**:

- z rozkładu $\mathcal{N}(x|0, 1)$,
- czy też z rozkładu $\mathcal{N}(x|1, 0.1)$.



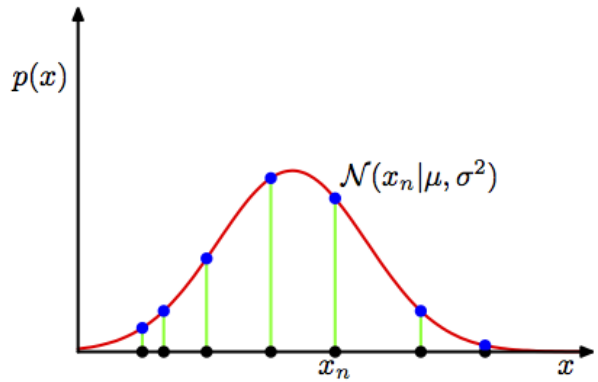
Wartości **funkcji wiarygodności**

$p(\mathcal{D}|0, 1) = 0.00037608,$
$p(\mathcal{D}|1, 0.1) = 2.15 \cdot 10^{-20}.$

Funkcja wiarygodności- Przykład 2

Dysponujemy ciągiem obserwacji
$\mathcal{D} = \{0.24, 0.32, 0.21, -0.2, -0.87,$
$0.23, -0.12, 0.01\}$

Chcemy ocenić, czy bardziej **wiarygodne** jest, że dane zostały **wygenerowane**:



- z rozkładu $\mathcal{N}(x|0, 1)$,
- czy też z rozkładu $\mathcal{N}(x|1, 0.1)$.

Wartości **funkcji wiarygodności**

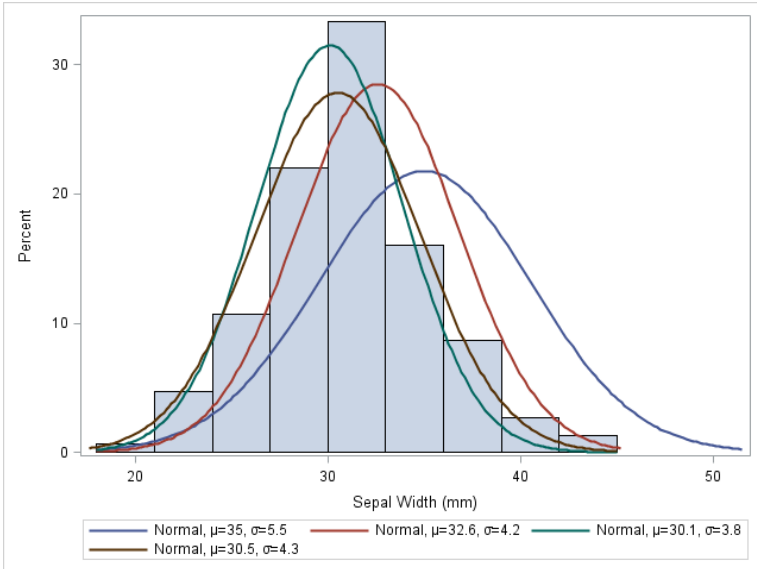
$p(\mathcal{D}|0, 1) = 0.00037608,$
$p(\mathcal{D}|1, 0.1) = 2.15 \cdot 10^{-20}.$

Estymator maksymalnej wiarygodności

Interesuje nas znalezienie modelu o **najwyższej wiarygodności**.

Formalnie, zadanie to formułujemy jako **zadanie optymalizacji**:

$$\begin{aligned}\boldsymbol{\theta}_{ML} &= \arg\max_{\boldsymbol{\theta}} p(\mathcal{D}|\boldsymbol{\theta}) \\ &= \arg\max_{\boldsymbol{\theta}} \prod_{n=1}^{N} p(\mathbf{x}_n|\boldsymbol{\theta}),\end{aligned}$$

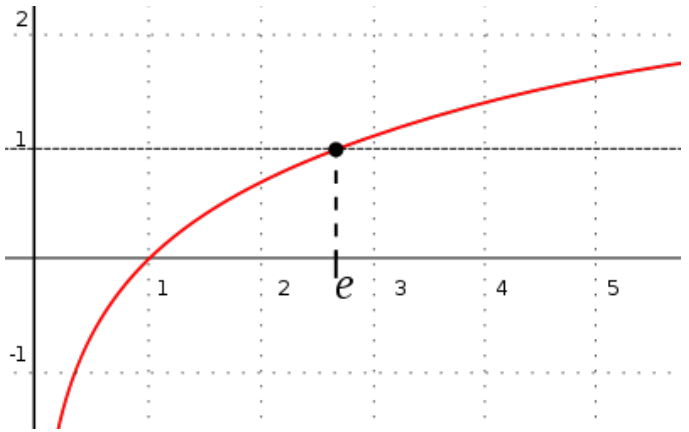


gdzie $\boldsymbol{\theta}_{ML}$ nazywany jest estymatorem maksymalnej wiarygodności (*ang. maximal likelihood estimate*).

Estymator maksymalnej wiarygodności

W praktyce definiuje się **alternatywne** zadanie **optymalizacji**:

$$\begin{aligned}\boldsymbol{\theta}_{ML} &= \arg\min_{\boldsymbol{\theta}} -\log p(\mathcal{D}|\boldsymbol{\theta}) \\ &= \arg\min_{\boldsymbol{\theta}} -\sum_{n=1}^{N} \log p(\mathbf{x}_n|\boldsymbol{\theta}).\end{aligned}$$



Okazuje się, że operacja logarytmu **nie zmienia** położenia **punktu optymalnego**!

Estymator ML - Przykład 1

Negatywny logarytm z **funkcji wiarygodności** dla rozkładu dwupunktowego:

$$
\begin{aligned}
-\log \mathrm{Bern}(\mathcal{D}|\theta) &= -\sum_{n=1}^{N} \{x_n \log \theta + (1 - x_n) \log(1 - \theta)\} \\
&= -m \log \theta - (N - m) \log(1 - \theta)
\end{aligned}
$$

gdzie m oznacza liczbę reszek.

Estymator ML:

$$
\theta_{ML} = \frac{m}{N}
$$

Estymator ML - Przykład 2

Negatywny logarytm z **funkcji wiarygodności** dla rozkładu normalnego:

$$-\log \mathcal{N}(\mathcal{D}|\mu, \sigma^2) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2 - \frac{N}{2}\log\sigma^2 - \frac{N}{2}\log 2\pi$$

Estymator ML parametru μ jest równy:

$$\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

Estymator ML parametru σ^2:

$$\sigma^2_{ML} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{ML})^2$$

Estymator maksymalnego *a posteriori*

- Załóżmy, że dysponujemy ciągiem obserwacji dotyczących rzutu monetą $\mathcal{D} = \{1, 1, 1, 1, 1\}$.

- Jeżeli wykonamy estymację parametrów **ML** parametru θ wówczas:

$$\theta_{ML} = \frac{m}{N} = \frac{5}{5} = 1.$$

- W rezultacie otrzymujemy rozkład **dwupunktowy** dla którego **prawdopodobieństwo** reszki wynosi $\mathbf{1}$.

- W celu rozwiązania tego problemu załóżmy, że **parametr** θ charakteryzuje się niepewnością.

Estymator maksymalnego *a posteriori*

Estymację formułujemy jako **zadanie optymalizacji**:

$$\boldsymbol{\theta}_{MAP} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathcal{D}),$$

gdzie $\boldsymbol{\theta}_{MAP}$ nazywany jest estymatorem maksymalnego a posteriori (*ang. maximal a posteriori estimate*).

Korzystając z reguły Bayesa mamy:

$$\boldsymbol{\theta}_{MAP} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta})p(\mathcal{D}|\boldsymbol{\theta}).$$

Estymator MAP - Przykład

Wprowadźmy rozkład na parametr θ w rozkładzie $\mathrm{Bern}(x|\theta)$:

$$Beta(\theta|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}.$$

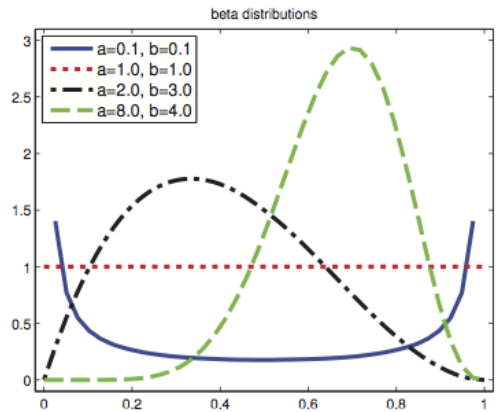
Wówczas mamy dla rozkładu *a posteriori*:

$$p(\theta|\mathcal{D}) \propto \frac{\Gamma(a+b+N)}{\Gamma(a+m)\Gamma(b+l)}\theta^{m+a-1}(1-\theta)^{l+b-1},$$

gdzie $l = N - m$.

Estymator **MAP**:

$$\theta_{MAP} = \frac{m+a-1}{N+a+b-2}$$



Modele generujące

- Modele **generujące** (*ang. generative models*) zakładają, że modelujemy łączne prawdopodobieństwo korzystając z **reguły łańcuchowej**:

$$p(x, y) = p(x|y)p(y)$$

- Do modelowania komponentów $p(y)$ i $p(x|y)$ wykorzystujemy elementarne modele i całość **składamy jak z klocków**.

Model GDA

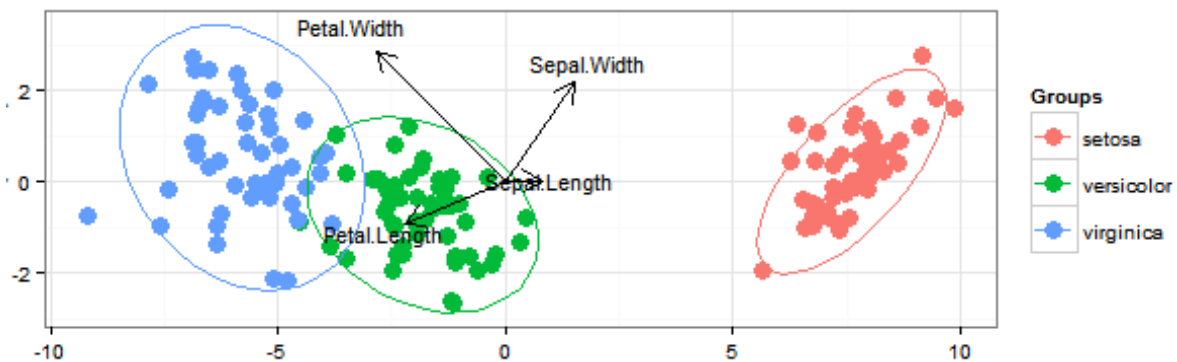
Model **GDA** (*ang. Gaussian Discriminant Analysis*) zakłada:

- Mamy K klas. Klasa reprezentowana jest w **schemacie one-hot** przez wektor $\mathbf{y}$ o rozkładzie:

$$p(\mathbf{y}) = \mathrm{Cat}(\mathbf{y}|\boldsymbol{\theta}) = \prod_{k=1}^{K} \theta_k^{y_k}$$

- Obserwacje $\mathbf{x}$ z każdej klasy modelowane przy pomocy wielowymiarowych rozkładów normalnych:

$$p(\mathbf{x}|\mathbf{y}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{y_k}$$



Model GDA

- Estymatory parametrów $\{\boldsymbol{\theta}, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K\}$ wyznaczamy metodą **największej wiarygodności**.

- Wnioskowanie, czyli wyznaczenie rozkładu *a posteriori* na klasę robimy ze **wzoru Bayesa**:

$$p(y_k = 1|\mathbf{x}) = \frac{\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\theta_k}{\sum_{j=1}^{K} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)\theta_j}$$

- **Predykujemy** poprzez wybór najbardziej prawdopodobnej klasy:

$$k^* = \arg\max_k p(y_k = 1|\mathbf{x})$$

Model Naive Bayes

- Model **Naiwnego Bayesa** (*ang. Naive Bayes*) zakłada **uproszczenie**, że poszczególne obserwacje z wektora $\mathbf{x}$ są **niezależne**:

$$p(\mathbf{x}|\mathbf{y}) = \prod_{d=1}^{D} p(x_d|\mathbf{y})$$

- W przypadku ciągłym $\mathbf{x} \in \mathbb{R}^D$, w modelu GDA używamy wtedy **diagonalnych macierzy kowariancji** postaci $\mathbf{\Sigma}_k = \mathrm{diag}(\boldsymbol{\sigma}_k)$:

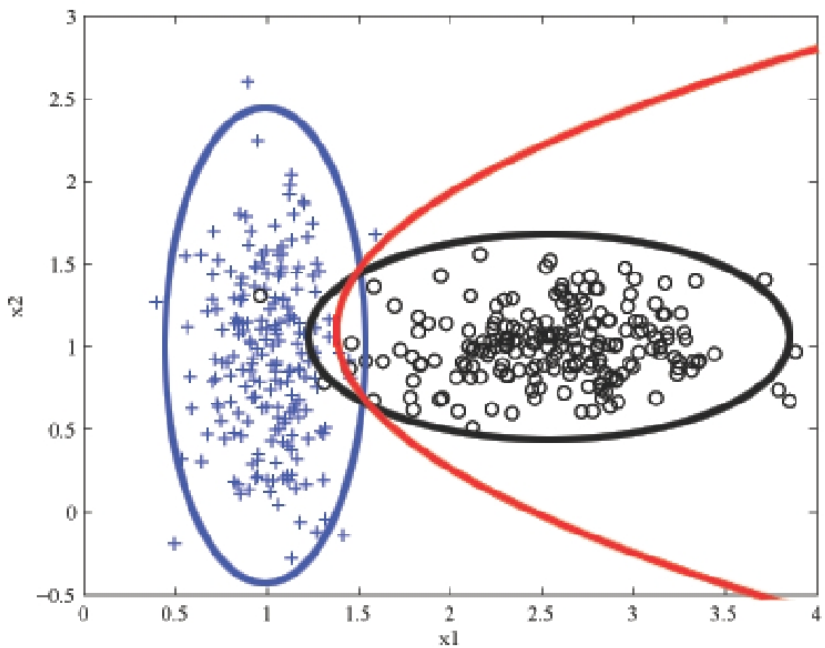
$$\begin{aligned} p(\mathbf{x}|\mathbf{y}) &= \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \mathrm{diag}(\boldsymbol{\sigma}_k))^{y_k} \\ &= \prod_{k=1}^{K} \prod_{d=1}^{D} \mathcal{N}(x_d|\mu_{k,d}, \sigma_{k,d})^{y_k} \end{aligned}$$

Model Naive Bayes

- Uproszczenie prowadzi do **istotnej redukcji** liczby parametrów:

$$\boldsymbol{\Sigma}_k \rightarrow \frac{D(D+1)}{2} \quad \text{VS.} \quad \text{diag}(\boldsymbol{\sigma}_k) \rightarrow D$$

- Jest to szczególnie istotne w przypadku danych **wysokowymiarowych**.



Model Naive Bayes

- Dla danych dyskretnych $\mathbf{x} \in \{0,1\}^D$ zakładamy, że cechy modelujemy **niezależnymi rozkładami dwupunktowymi**:

$$p(\mathbf{x}|\mathbf{y}) = \prod_{k=1}^{K} \prod_{d=1}^{D} \text{Bern}(x_d|\theta_{d,k})^{y_k}$$

- Założenie o niezależności jej tutaj jeszcze bardziej istotne ze względu na liczbę parametrów:

$$\text{rozkład pełny} \rightarrow 2^D - 1 \quad \text{VS.} \quad \text{niezależne cechy} \rightarrow D$$

- **Wnioskowanie** i **predykcja** są analogiczne jak w modelu GDA.